Turkish Journal of Health Science and Life

RELIABILITY ANALYSIS OF THE FEEDBACK SCALE OF A COURSE WITH CLASSICAL TEST THEORY AND GENERALIZABILITY THEORY

Giray KOLCU [a,b*], Mukadder İnci Başer KOLCU [a]

[a] Department of Medical Education and Informatics, Süleyman Demirel University, Isparta, Turkey

[b] Süleyman Demirel University Vice Director of Institute of Health Sciences, Isparta, Turkey

ARTICLE INFO

Article history:
Received: 08 August 2020
Accepted: 06 September 2020
Available Online: 25 October 2020

Key Words:
program evaluation, feedback, generalizbility theory, reliability

*Correspondence: Giray KOLCU ,
Department of Medical Education and Informatics, Süleyman Demirel University, Isparta, Turkey
e- mail: giraykolcu@gmail.com

Turkish Journal of Health Science and Life
2020, Vol.3, No.2, 20-24.

ABSTRACT

Introduction: The reliability of the measurement tools, preferred for the program evaluations for the development of pre-graduate medical education programs within the framework of the concept of accountability, should be high. (1–3) Reliability refers to the consistency of scores obtained with a specific measurement tool. (4,5).

Purpose: In our study, it is aimed to evaluate the internal consistency reliability analyzes of the qualitative research course feedback with classical test theory (Crohnbach's alpha) and generalizability theory (G-factor).

Method: In this study, the feedback data of 46 participants belonging to the qualitative research course applied by the researcher, were evaluated. Descriptive analysis of the data of the feedback scale and the reliability coefficients were determined according to the classical test theory and generalizability theory.

Results: In the evaluation of the feedbacks with the classical test theory, the Crohnbach's alpha coefficient was calculated as 0.947 for the evaluation of the 20-items measurement tool of 47 participants. In the evaluation of feedback with generalizability theory and in the estimation of variance components with one-sided crossed pattern, the variance for individuals was calculated as 0.30 and the estimated variance percentage was 41%, the variance for the items was calculated as 0.08, and the estimated variance percentage was 11%, the variance for the individual-item was calculated as 0.34 and the estimated variance percentage was 47%.

Discussion: Performing reliability analysis of measurement tools preferred as data collection tool for program evaluation is important for accountability. (1,2,6–8). 0.70 and above is accepted as the universal reliability standard in reliability analyzes in the literature. (9). The scale in our study was evaluated as "acceptable" according to the classical test theory and generalizability theory in the reliability analysis of the scale. As a result of the reliability analysis of the feedback scale evaluated within the scope of our study with two theories, we believe that it can be used as a highly reliable measurement tool in the evaluation of this training program.

1. Introduction

Accountability is mentioned among the basic features of medical schools in the modern medical education. The education program should be evaluated regularly and systematically for an accountable medical school (10). Medical schools should evaluate their educational programs with all their elements without digressing from the context of program evaluation. (7,8,11). Program evaluation is one of the important elements of the educational programs. (6,12). Various program evaluation models are used in the education programs. (3,12,13). Feedback is used as a data source in many of these models (1,3,13). The measurement tools preferred for the program evaluation should be reliable for the development of institutions within the framework of the concept of accountability (1–3). In this context, the reliability analysis of data collection tools is very important for the reliability of the program evaluation. (1,13,14).

Reliability refers to the consistency of the scores obtained with a specific measurement tool (4,5). Reliability in the field of psychometrics is grouped as internal consistency reliability, test-retest reliability, parallel forms reliability and inter-observer reliability. Internal consistency reliability methods are sorted as the split half method, Kuder-Richardson reliability coefficients, theta reliability coefficient, omega reliability coefficient, Guttman reliability coefficient and Cronbach's Alpha Reliability Coefficient. (15). The alpha coefficient method, developed by Cronbach (1951), is an internal consistency estimation method suitable for being used when the measuring tool is scored

between 1 and 5 in Likert-type scales (9,16).

Analyzes developed from many theories such as classical test theory and generalizability theory are used in the reliability of measurement tools (17). In classical test theory, reliability is expressed as the ratio of the actual score variance (systematic variance) to the observed score variance. This observed variance is equal to the sum of the actual score variance and the error variance (non-systematic variance) (18). The Cronbach's alpha coefficient is a weighted standard mean of change calculated by proportioning the total variances of the items in the scale to the general variance (16,19). While the sources of error cannot be evaluated in the classical test theory, it is possible to reach a single reliability value by considering more than one error source at the same time in the generalizability theory (20,21). Generalizability theory is a more extended version of the classical test theory from different aspects: it treats multiple sources of variance in a single analysis, allows the magnitude of each variance source to be determined and allows the calculation of the two different reliability coefficients (G-factor and phi coefficient respectively) for making both relative decisions based on the performance of individuals and absolute decisions about their performance (22,23) .

Purpose: In our study, it is aimed to evaluate the internal consistency reliability analyzes of the qualitative research course feedback with classical test theory (Crohnbach's alpha) and generalizability theory (G-factor).

Method: In this study, feedback data of 46 participants belonging to the "Basic Qualitative Research" course, which lasted for one day within the public health internship during the family medicine period in the pre-graduation medical education program of Süleyman Demirel University Faculty of Medicine, were evaluated. Permission was obtained from the participants to use their data. Feedback questions were developed by the research team. The assessment of the measurement tool was designed as a five-point Likert type scale (1-I do not agree at all, 2-I disagree, 3-Moderately agree, 4-I agree, 5-I completely agree) (24). In this context, all participants were asked to fill the feedback at the end of the training. SPSS package software and R Studio software were preferred for statistical analysis. (25).

In the descriptive data analysis of the answers, the distributions of the answers, the mean and standard deviations of the score equivalents were determined first. Then, reliability analysis was started. At this stage, the reliability coefficients were determined according to the classical test theory and the generalizability theory (Tables 1 and 2) (16,26).

Table 1. Cronbach's alpha coefficient formula

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K}\sigma_{Y_i}^2}{\sigma_X^2}\right)$$

Cronbach's alpha coefficient takes a value between 0 and 1, this value is accepted to be 0.70 and above (16). G-factor in the generalizability theory is interpreted as the reliability coefficient in the Classical Test Theory and it takes a value between 0 and 1 (26).

Findings:

In our study, the feedbacks of the Qualitative Research Course were evaluated. 47 participants who attended the course were asked to evaluate the feedbacks about the course (n: 47). The distributions of the participants' responses to the feedbacks and the

Table 2. Estimation formulas of variance components for one-sided crossed pattern [bxm]

Variance Source	Sum of squares	Sd	Mean of Squares	Estimated variance components
Participant (b)	SS_b	n_{b-1}	MS_b - SS_b / n_{b-1}	σ (^2b)
Item (m)	SS_m	n_{m-1}	MS_m - SS_m / n_{m-1}	σ (^2m)
bm,e	$SS_{bm,e}$	(n_{b-1}) (n_{m-1})	$MS_{bm,e}$- $SS_{bm,e}$/ $Sd_{bm,e}$	σ ($^2bm,e$)

mean of the evaluation score were calculated for the descriptive features of the data set (Table 3).

In the evaluation of the feedbacks with the classical test theory, the Crohnbach's alpha coefficient was calculated as 0.947 for the evaluation of the 20-item measurement tool of 47 participants (Table 4).

Table 4. Reliability analysis of feedbacks with the classical test theory

	Number of Items	Number of participants	Reliability Analysis
Feedback	20	47	0.947*
			0.842**
			0.835***

*Cronbach's Alpha
** Spearman Brown
***Guttmann Split Half

In the evaluation of feedback with generalizability theory and in the estimation of variance components with one-sided crossed pattern, the variance for individuals was calculated as 0.30 and the estimated variance percentage was 41%, the variance for the items was calculated as 0.08, and the estimated variance percentage was 11%, the variance for the individual-item was calculated as 0.34 and the estimated variance percentage was 47% (Table 5).

Discussion:

Program evaluation is among the important components of the educational programs. Data are collected from many sources for the program evaluation and these data are analyzed and transformed into the information. In medical

Table 3: Distribution of responses to feedback and mean scores

	I do not agree at all	I disagree	I moderately agree	I agree	I completely agree	Mean ± SS
Question 1			14.9%	36.2%	48.9%	4.34 ± 0.73
Question 2			4.3%	44.7%	51.1%	4.46 ± 0.58
Question 3			19.1%	40.4%	40.4%	4.21 ± 0.74
Question 4			12.8%	34.0%	53.2%	4.40 ± 0.71
Question 5	2.1%	6.4%	25.5%	29.8%	36.2%	3.91 ± 1.03
Question 6		2.1%	14.9%	42.6%	40.4%	4.21 ± 0.77
Question 7	4.3%	8.5 %	34.0%	27.7%	25.5%	3.61 ± 1.09
Question 8:	8.5 %	8.5 %	19.1%	31.9%	31.9%	3.70 ± 1.24
Question 9:		10.6%	23.4%	34.0%	31.9%	3.87 ± 0.99
Question 10		2.1%	10.6%	17.0%	70.2%	4.55 ± 0.77
Question 11:		2.1%	14.9%	19.1%	63.8%	4.44 ± 0.82
Question 12:			2.1%	19.1%	78.7%	4.76 ± 0.47
Question 13:			4.3%	36.2%	59.6%	4.55 ± 0.58
Question 14:			8.5 %	31.9%	59.6%	4.51 ± 0.65
Question 15:			12.8%	27.7%	59.6%	4.46 ± 0.71
Question 16:			14.9%	27.7%	57.4%	4.42 ± 0.74
Question 17:			10.6%	42.6%	46.8%	4.36 ± 0.67
Question 18:		4.3%	12.8%	36.2%	46.8%	4.25 ± 0.84
Question 19:			6.4%	36.2%	57.4%	4.51 ± 0.62
Question 20		2.1%	12.8%	25.5%	59.6%	4.42 ± 0.80

Table 5. Sources of variance, G and Phi coefficients of feedbacks

Variance Source	Df	SS	MS	Variance	Variance Percentage%
P	46	295.15	6.42	0.30	0.41
F1	19	81.42	4.29	0.08	0.11
P*F1	874	299.23	0.34	0.34	0.47
G factor: 0.947					
Phi coefficient: 0.934					
Absolute Error Variance = 0.021					

df Degree of freedom, SS sum of SS squares, MS mean of squares, SEM Standard error of measurement.
The observed variance ratio explained by the each surface is calculated by dividing the individual variance component by the total observed variance.

education, choosing reliable measurement tools in the context of program evaluation is very important in terms of accountability. Many theories are suggested in the reliability analysis of these measurement tool . (1,2,6–8). Among these theories, classical test theory comes to the fore due to its easy application and easy understanding of its background, while generalizability theory comes to the fore with its evaluation of many different error sources with a single analysis (27–30).

The feedback scale that we have developed for the program evaluation of the Qualitative Research Course was evaluated within the scope of our study. In the descriptive analysis of the scale, it was determined that all of the participants evaluated the scale and the scale represented a high level of satisfaction.

0.70 and above is accepted as the universal reliability standard in reliability analyzes in the literature (9). The scale in our study was evaluated as "acceptable" according to the classical test theory and generalizability theory in the reliability analysis. As the reliability coefficients of the feedback scale evaluated within the scope of our study are very close in the reliability analyzes performed with the classical test theory and the generalizability theory, we believe that this measurement tool can be used as a reliable measurement tool in the evaluation of this training program.

References

1. Medicine SU of. Undergraduate Medical Educaion Program Evaluation Strategy [Internet]. 2016 [cited 2019 Jan 4]. Available from: https://medicine.usask.ca/documents/ugme/EvaluationFramework.pdf
2. Walsh K. Evaluation in medical education [Internet]. [cited 2019 Jan 4]. Available from: https://www.monash.edu/__data/assets/pdf_file/0007/935692/evaluation-in-medical-education.pdf
3. Durning SJ, Hemmer P, Pangaro LN. The structure of program evaluation: An approach for evaluating a course, clerkship, or components of a residency or fellowship training program. Teach Learn Med. 2007;19(3):308–18.
4. Şencan H. Sosyal ve Davranışsal Ölçümlerde Güvenilirlik ve Geçerlilik (Validity and Reliability in Social and Behavioral Measurements). 2005.
5. Baykul Y. Eğitimde ve Psikolojide Ölçme: Klasik Test Teorisi ve Uygulaması. Ankara: ÖSYM Yayınları; 2000.
6. Vassar M, Wheeler DL, Davison M, Franklin J. Program Evaluation in Medical Education: An Overview of the Utilization-focused Approach Journal of Educational Evaluation for Health Professions. J Educ Eval Heal Prof [Internet]. 2010 [cited 2019 Jan 4];7:1. Available from: http://jeehp.org/
7. Francisco U of CS. Program Evaluation Policy (Medical Student) | UCSF Medical Education [Internet]. [cited 2019 Sep 7]. Available from: https://meded.ucsf.edu/policies-procedures/program-evaluation-policy-medical-student
8. Medicine DF of. Program Evaluation - Division of Medical Education - Dalhousie University [Internet]. [cited 2019 Sep 7]. Available from: https://medicine.dal.ca/departments/core-units/DME/education/programs-centres/program-evaluation.html
9. Tavakol M, Brennan RL. Medical education assessment: a brief overview of concepts in generalizability theory. Int J Med Educ. 2013;4(4):221–2.
10. Melle E Van, Flynn L, Kaba A, Tavares W, Horsley T. Program Evaluation for Competency-based medical Education: Are we making a difference? Int Conf Resid Educ. 2016;

11. Gandomkar R, Sandars J. The importance of context in medical education program evaluation. Med Teach [Internet]. 2018 Jan 2;40(1):106. Available from: https://doi.org/10.1080/0142159X.2017.1390215
12. Behçet P, Yrd O, Taha D. Eğitimde program geliştirme ve değerlendirme. Eğitimde program geliştirme ve değerlendirme. 2017.
13. Stufflebeam D. Evaluation models. New Dir Eval. 2001;2001 (89):7–98.
14. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. Adv Simul [Internet]. 2016;1 (1):1–12. Available from: http://dx.doi.org/10.1186/s41077-016-0033-y
15. Ercan İ, Kan İ. Ölçeklerde Güvenirlik ve Geçerlik. Uludağ Üniversitesi Tıp Fakültesi Derg. 2004;30(3):211–6.
16. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16:297–334.
17. Castanelli DJ, Moonen-van Loon JMW, Jolly B, Weller JM. The reliability of a portfolio of workplace-based assessments in anesthesia training. Can J Anesth Can d'anesthésie [Internet]. 2019;66(2):193–200. Available from: https://doi.org/10.1007/s12630-018-1251-7
18. Güler N. Genellenebilirlik Kuramı ve SPSS ile GENOVA Programlarıyla Hesaplanan G ve K Çalışmalarına İlişkin Sonuçların Karşılaştırılması Generalizability Theory and Comparison of the Results of G and D Studies Computed by SPSS and GENOVA Packet Programs. Eğitim ve Bilim. 2009;34 (154):93–104.
19. Cho E, Kim S. Cronbach's Coefficient Alpha: Well Known but Poorly Understood. Organ Res Methods. 2015;18(2):207–30.
20. Atilgan H. Genellenebilirlik kurami ve puanlayicilar arasi güvenirlik için örnek bir uygulama. Egit ve Bilim. 2005;(7):95–108.
21. Güler N, Eroğlu Y, Akbaba S. Reliability of Criterion-Dependent Measurement Tools According To Generalizability Theory: Application in the Case of Eating Skills. 2014;14:217–32.
22. Shavelson RJ, Webb NM. Generalizability theory: A primer. Generalizability theory: A primer. Thousand Oaks, CA, US: Sage Publications, Inc; 1991. xiii, 137–xiii, 137. (Measurement methods for the social sciences series, Vol. 1.).
23. Cronbach JL, Gleser GC NH& RN. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: John Wiley and Sons; 1972.
24. İbrahim Turan, Ümit Şimşek HA. The Use and Analysis of Likert Scales and Likert - Type Items in Educational Giriş. Sak Univ J Educ. 2015;(30):186–203.
25. Mushquash, C., & O'Connor BP. SPSS, SAS, and MATLAB programs for generalizability theory analyses. Behav Res Methods. 2006;38(3):542–7.
26. Brennan RL. Generalizability theory. Generalizability theory. New York, NY, US: Springer-Verlag Publishing; 2001. xx, 538–xx, 538. (Statistics for social science and public policy.).
27. Kassab SE, Du X, Toft E, Cyprian F, Al-Moslih A, Schmidt H, et al. Measuring medical students' professional competencies in a problem-based curriculum: a reliability study. BMC Med Educ [Internet]. 2019;19(1):155. Available from: https://doi.org/10.1186/s12909-019-1594-y
28. Bajwa NM, Nendaz MR, Galetto-Lacour A, Posfay-Barbe K, Yudkowsky R, Park YS. Can Professionalism Mini-Evaluation Exercise Scores Predict Medical Residency Performance? Validity Evidence Across Five Longitudinal Cohorts. Acad Med [Internet]. 9000;Publish Ah. Available from: https://journals.lww.com/academicmedicine/Fulltext/publishahead/Can_Professionalism_Mini_Evaluation_Exercise.97503.aspx
29. Monteiro S, Sullivan GM, Chan TM. Generalizability Theory Made Simple(r): An Introductory Primer to G-Studies. J Grad Med Educ [Internet]. 2019;11(4):365–70. Available from: http://www.jgme.org/doi/10.4300/JGME-D-19-00464.1
30. Taşdelen Teker G, Odabaşı O. Reliability of scores obtained from standardized patient and instructor assessments. Eur J Dent Educ [Internet]. 2019 May 1;23(2):88–94. Available from: https://doi.org/10.1111/eje.12406